DOCUMENT RESUME

ED 358 153 TM 019 952

AUTHOR	Wise, Steven L.; And Others
TITLE	An Investigation of Restricted Self-Adapted Testing.
PUB DATE	Apr 93
NOTE	13p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Atlanta, GA, April 13-15, 1993).
PUB TYPE	Reports - Research/Technical (143) -- Speeches/Conference Papers (150)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*Achievement Tests; *Adaptive Testing; Comparative Testing; *Computer Assisted Testing; *Difficulty Level; Elementary Education; *Elementary School Students; Mathematics Achievement; *Mathematics Tests; Pretests Posttests; Selection; Test Anxiety; Testing Problems; Test Items
IDENTIFIERS	Portland School District OR; *Restricted Self Adapted Testing

ABSTRACT

A new testing strategy that provides protection against the problem of having examinees in adaptive testing choose difficulty levels that are not matched to their proficiency levels was introduced and evaluated. The method, termed restricted self-adapted testing (RSAT), still provides examinees with a degree of control over the difficulty levels of their test items. The range of item choice is restricted to a region around the examinee's current proficiency estimate. Participants in this study were 186 students in grades 3 through 8 in the Portland (Oregon) Public School system during the winter of 1992-93, who were tested as part of an ongoing computerized adaptive testing program. Students were randomly assigned to a computerized adaptive test (CAT), a self-adaptive test (SADT), or RSAT in mathematics. Results indicate no differences between CAT and SADT conditions in terms of mean proficiency and mean posttest state anxiety. The basic RSAT method appears to hold promise for providing examinees with control over the testing situation, while preventing large mismatches between item difficulty choice and proficiency level. The RSAT procedure should be evaluated empirically. (SLD)

An Investigation of Restricted Self-Adapted Testing

Steven L. Wise

University of Nebraska-Lincoln

G. Gage Kingsbury & Ronald L. Houser

Portland Public Schools

Paper presented at the 1993 annual meeting of the National Council on Measurement in Education, Atlanta, GA

An Investigation of Restricted Self-Adapted Testing

Computerized adaptive testing (CAT) is an increasingly popular application of microcomputers in the estimation of examinee proficiency. Using a pool of items that have been calibrated using item response theory (IRT), a computer algorithm is employed to match the difficulty levels of the administered items to the estimated proficiency level of the examinee. This is an efficient testing process; according to the principles of IRT, the most informative items to administer to an examinee are those with difficulty parameters that are close to the examinee's proficiency level. At the beginning of a typical CAT session an examinee's proficiency estimate is quite unstable, with an accompanying variability in the difficulty levels of the items administered. Proficiency estimates tend to increase in stability, however, as more items are administered. This is reflected in an increased homogeneity in the difficulty levels of the items administered to a given examinee. The payoff in testing efficiency is usually substantial; markedly fewer items are needed per examinee in order to attain the same level of measurement precision as with a conventional test. Increased testing efficiency is the primary advantage of CAT.

Self-Adapted Testing

Efficiency is not, however, the only benefit that can be gained from the use of IRT in computer-based testing. Several years ago, Rocklin and O'Donnell (1987) explored an innovative application of IRT in computerized testing, termed *self-adapted testing*, in which the difficulty levels of the items administered are chosen by the examinee, rather than by a computer algorithm (as in a CAT). They found that examinees who received a self-adapted test (SAT) scored significantly higher (in terms of an IRT-based

proficiency estimate) than examinees receiving a conventional computerized test. Rocklin and O'Donnell interpreted the higher scores on the SAT as an indication that examinees were able to make effective and strategic choices among the items.

Rocklin and O'Donnell (1987) did not explicitly compare SAT and CAT tests. Such a comparison was made in Rocklin and O'Donnell (1991); they found that anxiety influences examinee performance less on a SAT than on a CAT. Wise, Plake, Johnson, and Roos (1992) compared the test performances of college-age examinees who were randomly assigned to take either a SAT or a CAT. They found that, relative to the CAT, examinees taking the SAT showed (a) significantly higher mean proficiency estimates and (b) significantly lower post-test state anxiety. An important difference between the Rocklin and O'Donnell (1991) and the Wise et al. (1992) studies was their use of item feedback. In the Rocklin and O'Donnell study, feedback was provided only on the SAT; in the Wise et al. study, feedback was provided on both test types. More recently, Vispoel and Coffman (in press) compared SAT and CAT versions of a music listening test for junior high school students, finding that (a) the SAT yielded higher mean estimated proficiency and (b) performance on the SAT was less influenced by test anxiety.

The importance of item feedback in a SAT was studied further by Roos, Plake, and Wise (1992). Using college-age examinees, Roos et al. compared SAT and CAT, with item feedback either present or absent. They found that the SAT yielded (a) a significantly higher mean proficiency estimate than the CAT, regardless of whether or not item feedback was given, and (b) significantly lower mean post-test state anxiety. Thus, the findings of Wise et al. (1992) were replicated and the mean score and anxiety differences between the SAT and the CAT were found even when item feedback was absent.

It is interesting to note that the differences that have been found between the mean proficiency estimates yielded by SAT and CAT are not easily explained within an IRT framework. According to the invariance property of IRT, there should be no mean differences in examinee performance between SAT and CAT. This property states that a given examinee's expected proficiency estimate will be the same regardless of which items are administered from the item pool. In the Wise et al. (1992) and the Roos et al. (1992) studies, examinees who were administered a SAT outperformed their randomly equivalent CAT counterparts. This suggests that the items were generally easier for examinees in the self-adapted condition, an apparent violation of IRT parameter invariance. This apparent inconsistency can be reconciled if one considers item parameters to be dependent on the context of test administration. That is, item parameters may be invariant across groups of examinees within a testing context, but not across contexts. It is likely that, when a SAT is used, extraneous psychological characteristics of examinees such as anxiety and motivation are changed, thus altering the testing context. This explanation recognizes that an examinee's success in passing a test item is not simply a function of ability, but is also influenced by psychological factors.

Why has a SAT shown an apparently positive effect on examinee test performance? The effect can be readily explained by the notion of *perceived control*. Numerous psychological research studies have shown that in a stressful situation (e.g., during a test), if people perceive that they have some control over the source of stress, they exhibit lower anxiety, increased motivation, and improved performance on cognitive tasks. Perlmuter and Monty (1977) provide an overview of this research.

The Costs of Self-Adapted Testing

There are costs, however, associated with the use of self-adapted testing. Wise et al. (1992) found that, compared to a CAT, a SAT (a) yielded significantly higher median standard error of proficiency estimation and (b) yielded a significantly longer median testing time. The items administered to examinees in the SAT condition were, on the average, not as informative as the items administered in the CAT condition. In the SAT condition, examinees were allowed to choose from a diverse set of difficulty levels. Examinees sometimes chose difficulty levels whose items were not well matched to their proficiency levels. These choices detracted from the efficiency of the SAT, relative to that of the CAT.

How often do examinees choose items that are not well matched to proficiency? For the 20-item tests used in Wise et al. (1992), the median standard errors of proficiency estimation for SAT and CAT were .37 and .35, respectively. Moreover, for the SAT they found a correlation of .68 between the difficulty level of the 15th item administered and the final proficiency estimate. These findings indicate that examinees taking the SAT tended to make item difficulty choices that were fairly well matched to proficiency. A few of the examinees who took the SAT, however, had very large standard errors. These examinees made poorly-matched item difficulty choices, resulting in inefficient proficiency estimation as indicated by the large standard errors.

The primary objective of this study is to introduce and evaluate a new testing strategy that provides protection against the problem of examinees choosing difficulty levels that are not matched to their proficiency levels, while still providing examinees a degree of control over the difficulty levels of their items. This testing procedure, termed a *restricted self-adapted test*

(RSAT), represents a hybrid of the SAT and CAT methods. In a typical SAT, before an item is administered, the examinee is allowed to choose among six difficulty levels that span the entire range of item difficulty in the pool. In an RSAT, the examinee is also allowed to choose among six difficulty levels, but the range of choice is restricted to a region around the examinee's current proficiency estimate. Thus, for example, a highly proficient examinee would choose among a set of difficulty levels that that are all comprised of the more difficult items, whereas a less proficient examinee would choose among difficulty levels containing less difficult items. The restricted choice characteristic of an RSAT should encourage examinee perceptions of control while preventing an examinee from making difficulty level choices that are too far away from his/her proficiency level.

The mean differences in proficiency level and anxiety found between SAT and CAT in the Wise et al. (1992) and the Roos et al. (1992) studies were based on very similar experimental procedures (i.e., same examinee population, item pool, setting). It is prudent, therefore, to investigate the generalizability of their results in different testing contexts. An additional objective of the present study was to compare SAT and CAT with younger examinees and a different item pool.

Method

Participants

Participants in this study were 186 students enrolled in grades 3 through 8 in the Portland Public School system during the winter of 1992-93. Students were tested as part of an ongoing computerized adaptive testing project, and were enrolled in four elementary schools, two middle schools, and three high schools. Each student was randomly assigned to one of the three testing conditions: CAT, SAT, or RSAT.

Instruments

Each student was administered a computer-based mathematics test. The test items were drawn from a pool of over 1000 multiple-choice items that had been previously calibrated using a one-parameter logistic IRT model. Each test was constrained to balance the content in eight goal areas, and was scored using a maximum-likelihood scoring procedure. Each test continued until the student had taken a maximum of 20 items, or until the standard error of the student's score dropped below .5 theta units. In all test conditions feedback was not provided after each item

In the CAT condition, items to be presented were selected to maximize information at the student's momentary proficiency estimate. To avoid extreme swings in difficulty at the beginning of the test, a Bayesian proficiency estimate was used for item selection purposes.

In the SAT condition, the student chose among five difficulty levels for each item (very easy, easy, average, hard, very hard). The algorithm then administered the item from the pool that maximized information at the chosen difficulty level. The five difficulty levels corresponded to -1.4, -0.4, 0.6, 1.6, 2.6 on the theta metric, respectively.

The RSAT condition was designed such that an examinee's five difficulty choices (a) were arranged around his/her proficiency estimate and (b) spanned only a portion of the theta scale. Once a proficiency estimate was calculated, the five difficulty level choices (labeled very easy to very hard) corresponded to the proficiency estimate value plus the following five adjustment values: -1.4, -0.6, 0.0, 0.6, and 1.4. For example, if an examinee had a proficiency estimate of 1.50 then the five difficulty level choices would correspond to: 0.1, 0.9, 1.5, 2.1, and 2.9. Once a choice value was determined, the algorithm would then administer the item from the pool that maximized

information at the choice value. Note that this highly proficient examinee would be prevented from choosing a very easy item, which would be uninformative from an IRT standpoint.

Unfortunately, due to a computer programming error, the RSAT procedure was not properly administered. This error was not recognized until very late in the data collection process. For this reason, data for the RSAT condition are not currently available. New data collection has begun, but data collection will not be completed for several months.

A shortened, 10-item version of the State Anxiety Scale of the State-Trait Anxiety Scale for Children (STAIC; Spielberger, 1973) was used to measure student post-test state anxiety. The scale was shortened to allow the total testing session to fit within a forty minute class period. The ten items used were those that had the highest item-total correlations reported in the STAIC user's manual. This scale was also computer administered.

Procedure

Each mathematics test was administered on a stand-alone PC-clone microcomputer by school personnel who were quite familiar with the test administration procedures. Students were given a very short keyboard familiarity exercise if they were unfamiliar with the testing system. Immediately after the mathematics test the shortened state scale was administered.

Data Analysis

Due to the computer programming error in the RSAT condition, only the data for the CAT and SAT conditions were analyzed. There were two dependent variables of interest: estimated proficiency and post-test state anxiety. The independent variable was test type (CAT, SAT). Two one-factor

analyses of covariance were performed, with grade level as the single covariate. A .05 level of significance was used in each analysis.

Results

Table 1 shows the treatment condition means and standard deviations for each dependent variable. The analysis of covariance for proficiency showed a nonsignificant treatment effect ($F(1,122) = 0.00$, $MS_E = 3.13$, $p = .993$). For post-test state anxiety, the treatment effect was also nonsignificant ($F(1,122) = 0.96$, $MS_E = 12.04$, $p = .330$).

Table 1

Descriptive Statistics, by Treatment Condition, for Proficiency and Post-Test State Anxiety

Dependent Variable	Test Type	Mean	SD	n
Proficiency	CAT	1.30	1.72	65
	SAT	1.29	1.92	60
Post-Test State Anxiety	CAT	17.49	3.58	65
	SAT	16.88	3.35	60

Discussion

The results indicated no difference between the CAT and SAT conditions in terms of mean proficiency and mean post-test state anxiety. In the present context, therefore, the differences found in previous studies (Wise

et. al., 1992; Roos et al., 1992; Vispoel & Coffman, in press) were not replicated. There are several potential explanations for these findings. First, the examinees may not have felt highly anxious about the test. If this were so, then the test situation may not have been perceived as being sufficiently stressful for the SAT to have the effects found in previous research. This explanation of low examinee anxiety does not appear to be supported by a comparison of the distribution of the state anxiety scores to the norms provided in the STAIC manual. The mean state anxiety score of approximately 17 found in this study for the shortened State Anxiety Scale corresponds to at least the 70th percentile in the normative sample. Moreover, inspection of the distribution of state anxiety scores revealed that many of the students reported substantial levels of anxiety.

A second explanation is developmental in nature. The majority of the students in this study were from the fourth or fifth grade. It is possible that, at these ages, the link between perceived control and stress reduction may not be highly developed. If this were the case, then providing students with control over their difficulty levels might not have an effect on anxiety and test performance. This issue is in need of additional research.

A third explanation, also developmental in nature, involves the lack of item feedback in this study's tests. Roos et al. (1992) found that, with a sample of adult examinees, item feedback was not necessary for the SAT to yield higher mean proficiency scores and lower mean anxiety than a CAT. Presumably, even in the absence of item feedback, an adult examinee is able to assess both the difficulty of a test item and the likelihood that he/she knew the correct answer. Younger examinees may not be able to make such assessments without item feedback being explicitly provided.

The effectiveness of the RSAT procedure has not yet been evaluated. In this study the basic RSAT method has been explained, and it appears to hold promise for providing examinees with control over the testing situation while preventing large mis-matches between item difficulty choice and proficiency level. It is important, however, that examinees perceive that the difficulty choices available to them are sufficiently diverse to engender feelings of control. If an examinee cannot discriminate between the difficulty levels of the items corresponding to the available choices then he/she will probably not perceive control. Hence, the RSAT procedure needs to be empirically evaluated. The data collection currently underway should provide useful information regarding the effectiveness of the RSAT procedure.

References

Perlmuter, L. C., & Monty, R. A. (1977). The importance of perceived control: Fact or fantasy? *American Scientist*, 65, 759-765.

Rocklin, T., & O'Donnell, A. M. (1987). Self-adapted testing: A performance-improving variant of computerized adaptive testing. *Journal of Educational Psychology*, *79*, 315-319.

Rocklin, T., & O'Donnell, A. M. (1991, April). *An empirical comparison of self adapted and maximum information item selection.* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Roos, L. L., Plake, B. S., & Wise, S. L. (1992, April). *The effects of feedback in computerized adaptive and self-adapted tests.* Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco.

Spielberger, C. D. (1973). *Manual for the State-Trait Anxiety Inventory for Children.* Palo Alto, CA: Consulting Psychologists Press.

Vispoel, W. P., & Coffman, D. D. (in press). Computerized adaptive and self-adapted music listening tests: psychometric features and motivational benefits. *Applied Measurement in Education.*

Wise, S. L., Plake, B. S., Johnson, P. L., & Roos, L. L. (1992). A comparison of self-adapted and computerized adaptive tests. *Journal of Educational Measurement*, 29(4), 329-339.